Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Environmental Sciences 12 (2012) 899 – 905

Procedia Environmental Sciences

2011 International Conference on Environmental Science and Engineering

(ICESE 2011)

Sampling Error in Ensemble Kalman Filter for 2-D Groundwater Flow*

Lin Lin, Lihua Xiong and Liangsheng Shi

State key laboratory of water resources and hydropower engineering science
Wuhan University
Wuhan, Hubei Province, 430072, China
linlin.wh@gmail.com

Abstract

Data assimilation is useful to determine key hydrogeological parameters, and EnKF proved to be an effective method. Sampling error exists in the EnKF if the sample size is not sufficiently large. This article tested the capability of the cubature-based EnKF based on the 2-D groundwater flow problem. The sensitivity of EnKF to Monte Carlo-based sampling techniques and the performance of the cubature-based EnKF were further analyzed. The results show that it is hard to decide which Monte Carlo-based sampling technique is with the least estimation variability when taking the same ensemble size. The cubature-based EnKF provides deterministic estimation result and effectively reduces the sampling error.

Keywords: Data assimilation; Sampling technique; Ensemble Kalman filter; Cubature rule

1.Introduction

The parameter estimation techniques and inverse modeling are used to determine key parameters affecting groundwater flow and solute transport in porous media. Data assimilation-based parameter estimation can be used to deterministically tune forecast models [1]. Ensemble Kalman Filter (EnKF) is a Monte Carlo-based data assimilation technique [2], which has been applied to investigate the groundwater flow and transport process in heterogeneous porous media [3]. The EnKF turns out to provide improved parameter estimation and history matching because consecutive measurements are used for continuously updating static parameters (porosity and hydraulic conductivity) and dynamic states (pressure head, flow velocity and solute concentration) [4]. The EnKF assumes that the covariance matrix can be approximately estimated from an ensemble of a series of sample. The cost of EnKF is significantly less than that of the traditional Kalman filter, since that EnKF avoids the rigid derivation of error covariance which is usually time-consuming in large-scale, non-linear systems. Because of the straightforwardness and efficiency of EnKF, it is also used extensively in oceanography [5], meteorology [6] and such large-scale applications.

However, considerable sampling error may be introduced in the EnKF if the sample size is not sufficiently large, especially when the sample size is small compared to the number of degrees of freedom (also called system dimension) [7]. As pointed out by them, sampling error consists of spurious correlation and estimation variability. The former

* This work is supported by NSFC Grant #51009110 to L. Shi, NSFC Grant #51079098 to L. Xiong and China Postdoctoral Science Foundation (Grant #20100471139) to L. Lin.

 doi:10.1016/j.proenv.2012.01.365

error describes the deterioration of forecast covariance when the rank of the covariance matrix is far larger than the sample size. The later error describes the phenomenon of that the assimilated parameters from different samples are usually fluctuated due to the sample variability in the covariance matrix. Most of the proposed sampling error reduction methods are based on variance reduction techniques [8]. Reference [9] introduced the deterministic ensemble determined by a proper numerical cubature rule technique to reduce the sampling error. Although there are a large amount of cubature rules, we emphasize that two conditions need to be satisfied to be used in the EnKF, i.e. the positive integration weights and equal weights. Both Gaussian and non-Gaussian problems can find the suitable cubature rules.

This research constructs horizontal two-dimensional groundwater numerical model, which is spatially discretized by Galerkin finite element method and takes Monte Carlo-based sampling method for the stochastic parameter of log saturated hydraulic conductivity. The capability of the cubature-based EnKF is testified by a numerical example. To compare the effect of EnKF and the cubature-based EnKF in reduction of ensemble sampling error, the sensitivity of EnKF to Monte Carlo-based sampling techniques and the performance of the cubature-based EnKF are analysed. In EnKF system, three algorithms are compared when forming initial ensemble stochastic field, i.e. Plain Monte Carlo (PMC), Quasi-Monte Carlo (QMC) and Latin hypercube sampling (LHS). In cubature-based EnKF system, the degree-2 and degree-3 rules are selected as equally weighted cubature rules based on Gaussian probability density function.

2.Flow equation and Ensemble Kalman Filter

According to Darcy's law and the continuity equation, the 2-D groundwater flow governing equation is as follows,

$$S_s \frac{\partial H(\mathbf{x},t)}{\partial t} = \frac{\partial}{\partial x_i} K_s(\mathbf{x}) \frac{\partial H(\mathbf{x},t)}{\partial x_j} - I(\mathbf{x},t) \tag{1}$$

Where, S_s is the specific storage, $H(\mathbf{x},t)$ is the total head, $K_s(\mathbf{x})$ is the hydraulic conductivity, $I(\mathbf{x},t)$ is sink or source item, x_i and x_j are spatial coordinates (i,j = 1,2). In this study, the hydraulic conductivity is considered as a log-normally distributed random space function and the specific storage is treated as a deterministic constant.

In this study, Monte Carlo method is used for solving the flow equation, which is processed on the basis of an ensemble of equally likely realizations. Deterministic flow equations are solved for each realization and the statistics of the flow quantities can be obtained through the ensemble.

The forecast model in EnKF is performed on each ensemble member independently:

$$S_k^{\mathrm{f}}(i) = F\left[S_k^{\mathrm{a}}(i-1)\right] + e_{1k}(i) \tag{2}$$

where, S is the state vector, which represents the state of the system, including model parameters, variables and other observations; F is a forecast operator, representing the flow equations for our study; k denotes the index of ensemble members; f and a indicate the forecast and assimilation procedure, respectively. The observation vector at the time step i for each ensemble member is give by:

$$d_k(i) = HS^t(i) + e_{2k}(i) \tag{3}$$

where, $d_k(i)$ denotes the observation vector; $HS^t(i)$ is the observation data obtained from the true field; $e_{1k}(i)$ and $e_{2k}(i)$ are independent white noises for forecast model and the observations, which are different between realizations.

The assimilation procedure starts with Kalman gain's calculation, i.e.

$$\left\langle S^{\mathrm{f}}(i) \right\rangle \approx \frac{1}{N_e} \sum_{k=1}^{N_e} S_k^{\mathrm{f}}(i) \tag{4}$$

$$P^{\mathrm{f}}(i) \approx \frac{1}{N_e - 1} \sum_{k=1}^{N_e} \left\{ \left[S_k^{\mathrm{f}}(i) - \left\langle S^{\mathrm{f}}(i) \right\rangle \right] \left[S_k^{\mathrm{f}}(i) - \left\langle S^{\mathrm{f}}(i) \right\rangle \right]^{\mathrm{T}} \right\} \tag{5}$$

$$K(i) = P^{\mathrm{f}}(i) H^{\mathrm{T}} \left[H P^{\mathrm{f}}(i) H^{\mathrm{T}} + R(i) \right]^{-1} \tag{6}$$

where, N_e denotes the total number of the ensemble members; P is the state error covariance matrix; K is the Kalman gain; H is the observation operator which represents the relationship between the state vector and the observation vector; R is the error covariance matrix of the observations; the superscript t stands for the true value.

Based on the Kalman gain, an updated ensemble of states and the new state error covariance $P^{\mathrm{a}}(i)$ can be computed as,

$$S_k^{\mathrm{a}}(i) = S_k^{\mathrm{f}}(i) + K(i)\left[d_k(i) - HS_k^{\mathrm{f}}(i)\right] \tag{7}$$

$$P^{\mathrm{a}}(i) = \left[I - K(i)H\right]P^{\mathrm{f}}(i) \tag{8}$$

In the latter procedure, the posterior mean and (co)variance of the state vector can be selectively calculated whenever and wherever needed, eliminating the needs of keeping track of the whole covariance matrix.

3.Sampling techniques

3.1.Sampling by Plain Monte Carlo algorithm

Plain (Crude) Monte Carlo (PMC) simply uses the definition of the mathematical expectation [10]. Let us consider the problem of the approximate computation of the integral $I = \int_G f(x)p(x)\mathrm{d}x$. Let ξ be a random point with probability density function $p(x)$. Introducing the random variable $\theta = f(\xi)$ with mathematical expectation equal to the value of the integral I, then $E(\theta) = \int_G f(x)p(x)\mathrm{d}x$. Let the random points $\xi_1, \xi_2, \cdots, \xi_N$ be independent realizations of the random point ξ with probability density function $p(x)$ and $\theta_1 = f(\xi_1), \cdots, \theta_N = f(\xi_N)$. Then an approximate value of I is

$$\overline{\theta}_N = \frac{1}{N}\sum_{i=1}^{N}\theta_i \tag{9}$$

Equation (9) defines the plain Monte Carlo algorithm: N is independent random point with probability density function $p(x)$ are generated. For each random point the value θ_i is computed. Then (9) is applied to get a Monte Carlo approximation to the value of I.

3.2.Sampling by Quasi-Monte Carlo approach

In Quasi-Monte Carlo (QMC) estimates [11], the nodes $\xi_1, \xi_2, \cdots, \xi_N$ are non-random points belonging to sequences uniformly distributed in I^n in the sense of Weyl. Such sequences having best asymptotic properties are called quasi-random (a somewhat misleading term). If the total variation of the integrand $f(x)$ is bounded, it follows from the Koksma-Hlawkain equality that as $N \to \infty$,

$$\delta_N = O\left(N^{-1}\log^n N\right) \tag{10}$$

Therefore, QMC is more efficient than PMC at all sufficiently large N. However, in practice the amount of points N is always restricted and the situation is not so clear. The most important problems where QMC is more efficient than PMC include integrands $f(x)$ whose dependence on x_i decreases as the number i is increased.

3.3.Sampling by Latin hypercube technique

Latin hypercube sampling (LHS) can be viewed as a compromise procedure that incorporates many of the desirable features of random sampling and stratified sampling and also produces more stable analysis outcomes than random sampling [12].

LHS operates in the following manner to generate a sample of size nS from $\mathrm{x} = [x_1, x_2, \cdots, x_{nX}]$ in consistency with the distributions $D_1, D_2, \cdots, D_{nX}$ (i.e. in consistency with the probability space). nX is the number of elements contained in x (i.e. $\mathrm{x} = [x_1, x_2, \cdots, x_{nX}]$). The range of each variable (i.e. the x_j) is exhaustively divided into nS

disjoint intervals of equal probability and one value is selected at random from each interval. The nS values thus obtained for x_1 are paired at random without replacement with the nS values obtained for x_2 : These nS pairs are combined in a random manner without replacement with the nS values of x_3 to form nS triples. This process is continued until a set of nS nX -tuples is formed. These nX -tuples are of the form and constitute the Latin hypercube sample. The individual x_j must be independent for the preceding construction procedure to work.

3.4.Deterministic sampling method

The cubature-based EnKF can reduce ensemble sampling error by providing deterministic ensemble determined by a proper equally weighted cubature rule [9]. With Gaussian probability density function, the degree-2 formula is a set of $N = n+1$ points,

$$(z)_k = (z_{k,1}, z_{k,2}, \cdots, z_{k,n}), \quad k = 0,1,\cdots,n \tag{11}$$

which are defined as

$$z_{k,2r-1} = \sqrt{2}\cos\frac{2rk\pi}{n+1}, z_{k,2r} = \sqrt{2}\sin\frac{2rk\pi}{n+1}, \quad r = 1,2,\cdots,[n/2],$$

where $[n/2]$ is the greatest integer less than $n/2$, and if n is odd, $z_{k,n} = (-1)^k$.

Similarly, a cubature rule of degree-3 is, $N = 2n$ equally weighted points of

$$(z)_k = (z_{k,1}, z_{k,2}, \cdots, z_{k,n}), \quad k = 1,2,\cdots,2n \tag{12}$$

with

$$z_{k,2r-1} = \sqrt{2}\cos\frac{(2r-1)k\pi}{n}, z_{k,2r} = \sqrt{2}\sin\frac{(2r-1)k\pi}{n}, \quad r = 1,2,\cdots,[n/2],$$

and if n is odd, $z_{k,n} = (-1)^k$.

In practice the performance of the above two rules is often better than the Monte Carlo method with much larger sampling points, even for relatively large dimensionality n.

4.Numerical example

A two-dimensional model of saturated flow is constructed to demonstrate the capability of the cubature-based EnKF by assimilating pressure heads and hydraulic conductivity measurements and to explore its efficiency in reducing sampling error compared with EnKF.

The flow equations are solved with a 2-D finite-element groundwater flow model. The flow domain is a square of size $L_x = L_y = 200[L]$ (L is any consistent length unit), uniformly discretized into 51×51 nodes. The grid is shown in Fig. 1, with 25 points denoting the observation locations. The left and right sides are prescribed as no-flow boundaries, and the upper and lower sides are the first-type boundaries with the constant head of 100[L]. There are an injection well at the location of (80, 80) with the flow of -5[L^2/T] and four pumping wells at the locations of (40, 40), (160, 40), (40, 160), (160, 160) with the same flow of 1[L^2/T] (where T is any consistent time unit).

The initial realizations of log hydraulic conductivity field for the cubature based EnKF system and EnKF system are generated by Karhuned-Loeve expansion [13] with a separable exponential covariance function, representing an initial guess to (prior knowledge of) the unknown true field. The mean and standard deviation of the initial realizations are taken as -2 and 1.0, respectively, and the integral scales are λ_x = 20.0[L] and λ_y = 20.0[L]. One realization of the PMC-based EnKF is randomly picked out, and is set as the reference field (the true underlying field), which gives us a log hydraulic conductivity field.

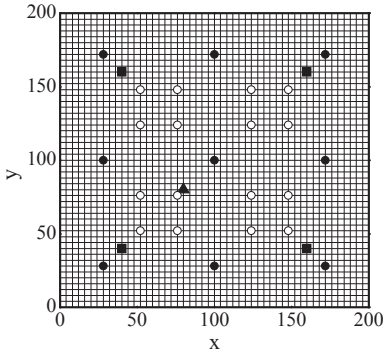


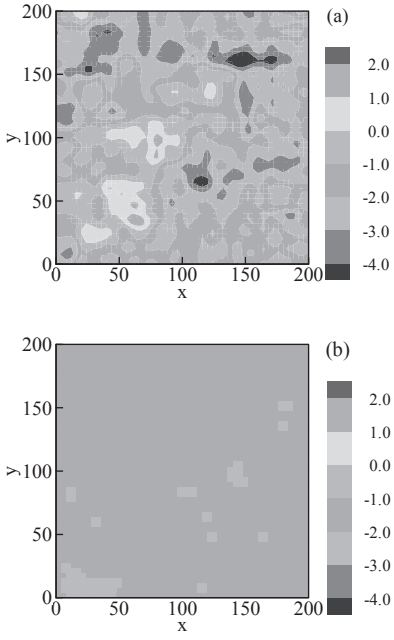
Fig. 1 The flow domain and observation locations. Pressure head observations are available at all the 25 circles; hydraulic conductivity measurements are available at the nine filled circles; the filled triangle is the injector and the filled squares are the producers.

At the measurement locations, the pressure head observations used in the two EnKF systems are drawn from the 2-D groundwater flow numerical simulation based on the reference field, and the $\ln K_s$ measurements are directly taken from the reference field.

In this model, the total simulation time is 1.0[T], which is evenly subdivided into 10 time steps of size 0.1[T]. Specific storage is assumed to be $1.0e^{-4}$[1/L]. The initial pressure head is taken as 100.0[L] throughout the domain. Since the same model (2-D groundwater flow) is used for solving both the forward step (forecast model) of EnKF and the reference pressure head field, we assume the system is free of model error.

5.Results

The capability of the cubature-based EnKF is testified by applying the numerical example introduced by the previous section. The reference field is with the mean equal to -1.750 and the standard deviation equal to 1.058 (Fig. 2(a)). The ensemble consists of 201 members (realizations) for the cubature rule of degree-2. For the cubature-based EnKF with the degree-2 rule the ensemble mean of the initial realizations and that of the estimated $\ln K_s$ field at the 10^{th} assimilation step are plotted in Fig. 2 (b) and (c). Compared with the reference $\ln K_s$ field, it is clear that the ensemble mean of the initial realizations does not show any decent features, although each realization represents the prior statistical knowledge of the reference field. However, after 10 assimilation steps, the ensemble mean field exhibits a pattern very similar to the reference field with the degree-2 and degree-3 rules. It is proved that the cubature-based EnKF is with the capability of recovering the major features of the reference field.



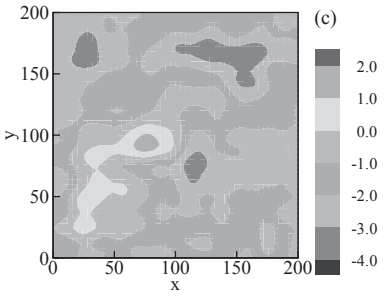


Fig. 2 Contours of the ln Ks fields in the example: (a) reference field; (b) ensemble mean of the initial realizations with the degree-2 formula; and (c) ensemble mean field at the 10th assimilation step with the degree-2 formula.

Since that the sampling technique is a necessary component of the initial realizations in the EnKF system, the sensitivity of EnKF to Monte Carlo sampling techniques should be analyzed. The initial ensemble consists of 200 and 400 members for the Monte Carlo-based EnKF. The ensemble consists of 201 and 400 members for the cubature rules of degree-2 and degree-3, respectively. The RMSE (root mean square error) is used to measure the goodness of the estimation matching the true field, which is calculated as,

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(S^{t} - S^{a}\right)^{2}} \tag{13}$$

For the three stochastic sampling techniques and each ensemble size (the number of the realizations) 17 series of RMSE are plotted in Fig. 3. Although it is hard to decide which Monte Carlo-based sampling technique is with the least estimation variability when taking the same ensemble size, the estimation variability of different series in EnKF system can be effectively avoided by taking the deterministic sampling strategy.

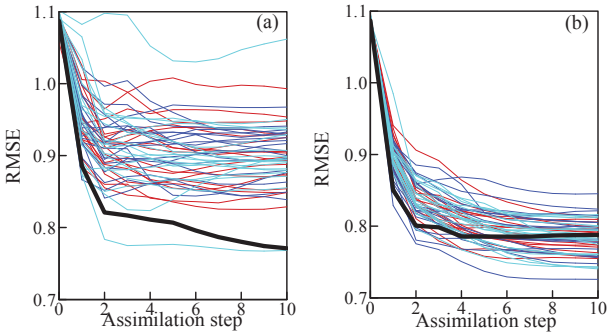


Fig. 3 Influence of sampling techniques: (a) 200 realizations; (b) 400 realizations. The black lines in (a) and (b) represent the cubature-based EnKF with the degree-2 rule and the degree-3 rule, respectively. The red lines, blue lines and cyan lines are PMC, QMC and LHS, respectively.

6.Conclusion

In this study, ensemble Kalman filter (EnKF) and the cubature-based EnKF are applied for continuously inversing the hydrogeological parameter. The two cubature-based EnKF methods prove to generate much more stable and accurate results than the three Monte Carlo-based EnKF.

References

[1] J.A. Hansen, and C. Penland, “On stochastic parameter estimation using data assimilation,”. *Physica D*, vol. 230, pp. 88-98, 2007.

[2] G.Evensen, “Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics,” *J. Geophys. Res*. Vol. 99, pp. 10143-10162, 1994.

[3] Y. Chen, D. Zhang, “Data assimilation for transient flow in geologic formations via ensemble Kalman filter,” *Adv. Water Res.,* vol. 29, pp. 1107–1122, 2006.

[4] A. Bianco, A. Cominelli, L. Dovera, G. Naevdal, and B. Valles, "History matching and production forecast uncertainty by means of the ensemble Kalman filter: a real field application SPE 107161," 2007.

[5] Z. Deng, Y. Tang, and G. Wang, "Assimilation of Argo temperature and salinity profiles using a bias-aware localized EnKF system for the Pacific Ocean," *Ocean Modelling*, vol. 35, no.3, pp. 187-205, 2010.

[6] J.C. Hargreaves, and J.D. Annan, "Using ensemble prediction methods to examine regional climate variation under global warming scenarios," *Ocean Modelling*, vol. 11, no. 1-2, pp. 174-192, 2006.

[7] R. Furrer, and T. Bengtsson, "Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants," *J. Multivariate Ana.*, vol. 98, pp. 227-255, 2007.

[8] G. Evensen, "Sampling strategies and square root analysis schemes for the EnKF," *Ocean Dyn.,* vol. 54, pp. 539-560, 2004.

[9] J. Li, and D. Xiu, "On numerical properties of the ensemble Kalman filter for data assimilation," *Computer Methods in Applied Mechanics and Engineering*. Vol. 197, pp. 3574-3583, 2008.

[10] I. Dimov, and R. Georgieva, "Monte Carlo algorithms for evaluating Sobol' sensitivity indices," *Mathematics and Computers in Simulation*, doi:10.1016/j.matcom.2009.09.005.

[11] I.M. Sobol', and D.I. Asotsky, "One more experiment on estimating high-dimensional integrals by quasi-Monte Carlo methods," *Mathematics and Computers in Simulation*, vol. 62, no. 3-6, pp. 255-263, 2003.

[12] J.C. Helton, and F.J. Davis, "Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems," *Reliability Engineering and System Safety*, vol. 81, pp. 23-69, 2003.

[13] D. Zhang, and Z. Lu, "An efficient, high-order perturbation approach for flow in random porous media via Karhunen–Loeve and polynomial expansions," *J. Comput. Phys.*, vol. 194, pp. 773-794, 2004.